

## REMARKS/ARGUMENTS

1. The amendments made in this response to the Office Action are intended to:
  - (A) inactivate the hyperlinks that were objected to in paragraph 3, page 2 of the Office Action,
  - (B) make clear that the instant invention does *not* perform text classifications or make use of training/test sets, even though classification and the use of training sets was contemplated in claims that had been withdrawn in connection with the election required by the Office Action of April 22, 2003 (i.e., the instant invention does not make use of most of the features of the Rainbow computer program.),
  - (C) elaborate on prior art concerning the association of genes with a text corpus of bibliographic literature;
  - (D) edit the claims so as to make more clear the scope of the patent claims.
2. The present amendments to the specification Nos. 3-15 inactivate hyperlinks that were objected to in **page 2, paragraph 3 of the Office Action**, specifically those on page 17, line 7 and page 18, line 14. The hyperlinks were inactivated by removing “http://” or “ftp://” from the text, which are now implied to the reader, but which should now not be interpreted as a hyperlink by a Web browser.
3. Additional amendments to the specifications were made (Nos. 16-19) in consideration of potential sources of confusion or misunderstanding, and in response to a reading of the Office Action. No new matter was introduced by these amendments.  
Amendments No. 16-18 were made in order to make clear that the instant invention does *not* perform text classifications and therefore does not make use of training/test sets, even though classification and the use of training sets were contemplated in claims that had been withdrawn, as

well as contemplated in matter that had been previously edited out in the connection with the earlier office action. Thus, in Applicant's October 20, 2003 Response to the Office Action of April 22, 2003, the revised Fig. 1 eliminated computer program modules beyond the Text Modeling Module (126) and Keyword Identification Module (128). In particular, the text classification module was eliminated from Fig. 1. Furthermore, the specification related to the text classification module was eliminated in paragraph 30 of the October 20, 2003 Amendments to the Specification, including all specifications pertaining to training and testing sets.

Amendment No. 19 expands the discussion of the prior art concerning the generation of a literature text corpus.

4. In the Office Action, Claim 1 was rejected under 35 U.S.C. 103(a) as being un-patentable under Andrade et al. (1999) taken with McCallum (1998) because it would have been obvious to one having ordinary skill in the art at the time of the invention to make an automated genome sequence analysis and annotation system to have the computer program, Rainbow, to perform the text and analysis classification, as taught by Andrade and McCallum.

Applicant thanks the Examiner for calling his attention to Andrade et al. (1999). However, for reasons described below, Applicant believes that Claim 1 is patentable over Andrade et al. (1999) taken with McCallum (1998), and that the amendments in this Office Action response will assist the Examiner in interpreting and justifying the scope of the patent claim. Applicant also believes that a consideration of the modifications needed to be applied to the references, in order to arrive at the claimed subject matter, are such that it would not have been obvious to one having ordinary skill in the art at the time of the invention to have made the invention. The accompanying **Declaration of David R. Rigney** is intended to assist the Examiner in considering who would be an artisan having ordinary skill in the art at the time of the invention.

## 5. THE SCOPE AND CONTENTS OF THE CITED PRIOR ART

5A. Before discussing the differences between the claims for Applicant's invention, versus the prior art taught in Andrade et al (1999) taken with McCallum (1998), the paragraphs that follow summarize the scope and contents of Andrade et al., which describes the GeneQuiz software.

5B. "GeneQuiz is a semi-automated protein sequence analysis system, the principal purpose of which is to infer a specific and reliable functional assignment by analysis of annotations from sequence database matches" (p. 392, column 1, line 38).

5C. "The GeneQuiz system is designed to analyze a single query protein sequence, or a batch of sequences such as a set of translated ORFs from a sequencing project ..." (page 400, column 1, line 45).

5D. "A single GeneQuiz run is triggered by entry of a query protein sequence into the system, or as one of a batch of sequences, perhaps representing the protein set of a full genome. ... The query is screened against the non-redundant sequence databases using several standard database search programs, and a multiple alignment [of protein sequences] is constructed." (p. 393, column 1, line 11).

5E. The "non-redundant sequence databases" against which the query protein sequence is screened are shown in Table 1 (p. 394). The "standard database search programs" are shown in Table 2 (p. 395), in particular the BLAST and FASTA tools that are used to score the similarity of matches between the query sequence and sequence database members, as shown in Table 5 (p. 398).

5F. "The GeneQuiz system takes as input a protein sequence and produces as output a specific functional annotation and general functional class for this sequence" (p. 392, column 2, line 47, with added emphasis to indicate what is input and what is output).

The input "protein sequence" is well understood in the art to mean a string of amino acid symbols, such as "VGHNADLQIKLSIRRLLAAGVLKQTKGVGAS....". By "general functional class for this sequence", Andrade et al. mean one of the 16 functional classes shown in Table 4 (p. 398).

By “specific functional annotation for this sequence”, Andrade et al. mean the types of words or phrases that are shown in Table 3 (p. 397).

5G. Output are generated in a GeneQuiz module called GQreason (p. 396, column 1, line 31):

“The main purposes of the GQreason module are 2-fold: (i) to determine a broad cellular function for the query sequence family by analyzing the set of homologues to the protein [that were found through sequence database searches using sequence comparison tools such as BLAST and FASTA], i.e., to assign the family to a general functional class; (ii) to assign a specific function to the query, if possible, by transferring that function from one of the homologues. Both tasks depend on the careful choice of homologues and on the systematic analysis of sequence database annotations”(added underlined emphasis of the word transferring, which is distinguished from an *ab initio* assignment of function).

“The GQreason module selects a set of sequences [that are] similar to the query sequence as reliable homologues. If none [of the homologues] has a characterized function, the module cannot assign a function to the query.” (page 403, column 1, line 37)

5H. As indicated in the previous cited text, the annotations that are to be transferred to the query sequence already exist in the sequence databases, and are found attached in those sequence databases to the homologues that match the query sequence. As taught by Andrade et al., those database annotations are created by expert human annotators using vaguely-defined criteria, not by the well-defined algorithm of an automatic natural language processing computer program:

“Systematic extraction of functional information from [pre-existing] annotations expressed in various database-specific field types (description, keyword, comment, etc.) and formats presents a harder problem”. (p. 396, column 1, line 45). ... “The annotations currently found in databases are highly heterogeneous and sometimes inconsistent in the use of database fields. ... Annotations are generally hand crafted and inevitably reflect idiosyncrasies of the [human] annotator despite attempts at standardization by the [human] curators. Typical forms of description include those shown in Table 3”. (p. 396, column 2, line 23 with added text between brackets for clarification).

5I. “The method used by GeneQuiz for general functional classification is based on generation of a dictionary that associates keywords characteristic of a sequence with a set of functional classes. ... GeneQuiz currently works with 14 classes of cellular function ... (Table 4) ...” (page 397, column 1, line 6).

5J. “Assignment of a new sequence to a class is by look-up of the keywords for that sequence in the dictionary to determine the most frequently associated class, which is then chosen” (page 397, column 2, line 13). “In GeneQuiz, the keywords associated with the majority of all SWISS-

PROT homologues of the query sequence – having suitable keyword information – are selected. Then the dictionary of keyword/class associations is used to attempt a classification of the query into one of the 14 functional classes.” (page 398, column 1, line 4).

5K. “GeneQuiz applies a lexical analysis procedure to the description fields of the query homologues to recognize likely functionally meaningful annotations.” (page 398, column 2, line 5). ... “The GQreason module applies the following algorithmic approach to the list of homologues: (i) for each database search method, assemble a separate list of homologues, descriptions and scoring information ordered by similarity to the query; (ii) transpose method-specific scoring into a common ‘reliability value’ scheme which incorporates biases favoring certain databases ... as detailed in Table 5; (iii) concatenate the lists placing favored search methods first (BLAST>FASTA); (iv) iterate over the partially sorted list, applying a lexical analysis to each functional description, either accepting or rejecting it according to the forms shown in Table 3. Lexical analysis consists of a series of tests for informational content ... Refer .. to Figure 2 ... Of the descriptions accepted (if any), the one having the highest reliability value [see Table 5] is carried over as the functional annotation of the query sequence, and the procedure terminates.” (page 398, column 2, line 13).

5L. “The user accesses the results of a GeneQuiz analysis via a set of HTML pages containing tabular information and graphical displays of alignments and structures, that is navigable in any table-compliant Web browser”(page 398, column 2, line 45). ... Findings for each ORF [open reading frame within a whole-genome ORF set] ... or for a user-supplied protein query submitted to the server, are presented in the form of a report giving structured access to and views of the collected results ... (page 399, column 1, line 4).

AGREEMENTS AND OBJECTIONS TO CHARACTERIZATIONS IN THE OFFICE ACTION REGARDING THE SCOPE AND CONTENTS OF THE CITED PRIOR ART

6. On page 2, paragraph 6 of the Office Action, the Examiner states “Andrade et al. discloses a system, GeneQuiz, for automated genome sequence analysis and annotation wherein a Web-based browser provides views of results, and links to biological databases (Abstract etc.) such as GenBank (page 396, column 2, Database Quality) and PROSITE (Table 1)” (underlining of Abstract etc. added for emphasis). Applicant agrees that GeneQuiz links to the databases shown in Table 1 of Andrade et al (page 394). However, Applicant disputes the suggestion that these databases contain anything that resembles “Abstracts” or that GeneQuiz links to biological databases that contain “Abstracts” or that Andrade’s description of GeneQuiz teaches that such “Abstracts” could be used by GeneQuiz even if they were linked to GeneQuiz. The databases

shown in Table 1, to which GeneQuiz links, such as Genbank and PROSITE, are databases of sequences (nucleic acid and protein). They are not literature databases, such as MEDLINE/Entrez PubMed, which do contain literature “Abstracts”. Applicant would agree with the Examiner’s characterization of GeneQuiz if the Office Action paragraph 6 were to read “...links to biological sequence databases, such as Genbank and PROSITE”, to make clear that the linkage is to sequence databases and that these databases do not contain literature abstracts, and that even if they did contain literature abstracts, such abstracts would not be used by GeneQuiz as taught by Andrade et al.

7A. On page 3, paragraph 7 of the Office Action, the Examiner states that “The system of Andrade et al. has been demonstrated with several gene sets of several genomes, which identify a new function, a new family, and a new superfamily (page 403, column1, New Findings).” Applicant agrees that the sentence is factually correct but disputes inferences that might be drawn from the sentence, in particular that these are “New Findings” or that they are relevant to Applicant’s claims. Although the GeneQuiz system has been demonstrated with sets of genomes, each containing of a set of translated ORFs from a sequencing project, it does not follow from this fact that GeneQuiz analyzes the translated ORFs simultaneously. Instead, Andrade et al teaches that the translated ORFs are analyzed one-by-one, without reference to one another, as a batch process (see paragraphs 5C and 5D above). It is well understood by those experienced in the art of computer programming that the term “batch” refers to the practice of placing a list of queries or instructions in a file, and submitting them for execution, so that a computer program can execute them one after the other, as though the queries or instructions were being typed manually and sequentially by a computer operator, and in such a way that a new query or instruction is not executed until execution of a previous one has been completed. In other words, GeneQuiz does

not analyze groups of sequences, except in the trivial sense that it permits its user make an individual sequence query, and upon completion of the analysis of that query, GeneQuiz permits the user to make another individual sequence query.

7B. The “new function” referred to by Andrade et al. on page 403, column 1, line 10, refers to the fact that the GeneQuiz results obtained in the year 1997 were different from the results that would have been generated in the year 1994, because the external sequence databases used by GeneQuiz to transfer annotations had been updated from 1994 to 1997. The fact that a “new function” could be inferred by GeneQuiz is not a new finding, because Andrade et al. teach that “Other non-automatic systems for protein comparison arrived at the same conclusions” (page 403, column 1, line 31), and it is not clear which system noticed the new function first. Presumably, this would simply be a matter of being the first to use the updated external sequence databases. Applicant would, however, agree that the GeneQuiz system is useful in this context because GeneQuiz was able to achieve the result in an automatic, labor-saving fashion.

7C. The “new family” referred to by Andrade et al. on page 403, column 1, line 37, pertains to a situation in which the homologues corresponding to a query sequence have no functional annotations in the database, so that GeneQuiz cannot assign any function or annotation to that query sequence. It is therefore a situation in which GeneQuiz fails in its intended goal of providing annotations for a query sequence. The utility of GeneQuiz in this situation is that it aids in the recognition that a family of previously un-annotated sequences exist, the members of which are defined in terms of their sequence similarity. However, this utility is not relevant to Applicant’s claims because Applicant’s claims deal with methods for annotating groups of genes *ab initio*, in which the sets of genes are always un-annotated prior to the use of Applicant’s methods.

Furthermore, Applicant's methods have nothing to do with sequence similarities among the sets of genes that are to be annotated.

7D. The "new superfamily" referred to by Andrade et al on page 403, column 2, line 44, pertains to situations in which the functional assignments made by GeneQuiz are unreliable by its own reckoning. Andrade et al teach that under these circumstances, "The system cannot resolve such cases automatically because of the risk of introducing many false positives into the putative family. However, the user, if particularly interested in a specific protein, has available all of the information derived by GeneQuiz and can manually validate candidate homologies..." (page 403, column 2, line 51). Thus, under these circumstances, it is the human expert who must manually discover new results by criteria that are unspecified by Andrade et al. If a by "new finding" the Examiner means a finding that is made by GeneQuiz without the participation of a human expert, then this is not a new finding. However, Applicant would agree that in this situation, GeneQuiz is useful, not because it provides new findings, but because it aids in pointing a human expert to where new findings might be discovered manually.

7E. On page 3, paragraph 7 of the Office Action, the Examiner states that "Further, Andrade et al. discloses improvements with said system to cluster database sequences into families with pre-processed functional annotation (page 409, column 1, 42-45), as in the instant steps (a) and (b)." Applicant disputes that the teachings are "as in the instant steps (a) and (b)". By "cluster the database sequences into families", Andrade et al. mean that the objects that are being organized into families are sequences, and the criteria by which the sequences are to be organized into families is on the basis of sequence similarity among members of the family. No other interpretation would be reasonable because GeneQuiz is intended to match a query sequence with sequences in sequence databases on the basis of sequence similarity. The instant

steps (a) and (b), on the other hand, are concerned with genes, but a sequence does not necessarily correspond to a gene, and even if a gene corresponded to a sequence it may be assigned to more than one sequence families as taught by Andrade et al. This is because Andrade et al. are concerned with the sequences of complete genomes that are found in sequence databases (see for example, Table 6, p. 402), and it is well understood by those skilled in bioinformatics and molecular biology that only about 5% of the sequences in the genome correspond to genes. Furthermore, in the sentence cited by the Examiner pertaining to the placement of sequences into families (page 409, column 1, 42-45), Andrade et al teach that this is to be as approached by the methods disclosed by Sonnhammer et al (1997). There is nothing in the approach of Sonnhammer et al (1997) to suggest that a sequence must be assigned to a single family, and in fact Sonnhammer et al provide evidence to the contrary. Sonnhammer et al discuss the “Pfam” software, which is concerned with the identification of *domain* families (domain = short subsequences within larger sequences), and there is no reason to expect that a sequence must contain only a single domain, as indicated in Sonnhammer et al. (page 413, column 1, line 11):

To what extent are proteins modular? With Pfam, we can address this problem with higher accuracy than before. Of the proteins in Swissprot 33 containing at least one Pham-A domain, 17% contain two or more domains, whereas 2.5% have five or more domains. This is only a lower bound because: 1) not all domains are present in Pfam-A, 2)HMMs [hidden Markov models] are not perfectly sensitive, and 3) it is based on proteins in Swissprot, which probably is biased towards single domain proteins. We have done the same analysis with Wormpep 10, which should represent a relatively unbiased set of proteins. Twenty-eight percent of the proteins that matched Pfam-A families matched in two or more domains, whereas 4% matched in five or more domains.

Consequently, according to Andrade et al., a sequence (which may or may not correspond to a gene) may be assigned to more than one family. Accordingly, a set of sequences (which may or may not correspond to a set of genes) would not in general be partitioned into subsets corresponding to the families contemplated by Andrade, simply because a sequence may contain

more than one sequence domain. Therefore, Andrade et al does not teach as in the instant steps (a) and (b), because Andrade et al. does not teach that a set should be partitioned disjointly.

**7F.** In the sentence cited by the Examiner on page 3, paragraph 7 of the Office Action, Andrade et al use the term “cluster” to mean something different than what is meant in the instant (b), namely, to partition a set into disjoint subsets. Applicant believes that Applicant’s use of this term is standard, corresponding usage of the term “cluster” in prior art cited in Applicant’s original Information Disclosure Statement. Andrade et al. use the term instead to mean the identification of similar properties within a collection of objects, without the partitioning of those objects into mutually exclusive subsets. Andrade et al. introduce this topic in connection with a discussion entitled “Scalability”, which is concerned with the problem of how to make GeneQuiz produce results within a reasonable period of time, given the fact that the sequence databases that it uses are becoming too large to make their use practical (page 409, column 1, line 16). It appears to Applicant that Andrade et al. are suggesting an improvement wherein prior to their use with GeneQuiz, the sequence databases shown in Table 1 (page 394) could pre-processed by software like Pfam (Sonnhammer et al, 1997) to identify domains within the database sequences, and that for reasons that must be guessed by the reader, GeneQuiz would run faster using the pre-processed sequence databases. Andrade et al suggest this approach as an alternative to the approach that was suggested in the previous paragraph (page 409, column 1, para. 4), which is to filter the databases for non-redundancy at a level lower than that presently used in GeneQuiz. Therefore, Andrade et al might be suggesting, in the sentence cited in the Office Action, that in order to make GeneQuiz run faster, the sequence databases could be filtered by eliminating from consideration those sequences that do not contain a domain that is recognized by Pfam, or that if

a query sequence contains a domain that is recognized by Pfam, then GeneQuiz would run faster by attempting to match that query sequence only with sequences in the database that also contain that domain. In either case, the suggestion would be unrelated to the claim of partitioning a set of input genes into subsets as in the instant (a) and (b), because the suggestion of Andrade et al. has to do with the sequence databases used by GeneQuiz and not with the (individually submitted or individual within a batch) query sequence that constitutes the input to GeneQuiz.

7G. The same objection -- that the improvements that Andrade et al suggest on page 409 in the section entitled "Scalability" do not pertain to the partitioning of a set of input genes as in the instant (a) and (b) -- would apply to any other hypothetical method for pre-processing of the sequence databases that are used by Andrade et al (as distinguished from the pre-processing of input query sequences), even if the hypothetical method did involve the partitioning of a set into disjoint subsets. Applicant could devise such a hypothetical method as follows. In the article Benson et al. concerning GenBank, which was attached to the Office Action, it is stated on page 2, column 2, line 51:

In order to organize the [GenBank sequence] data in a more useful fashion, NCBI has created the UniGene collection of unique human genes. ... Briefly, UniGene starts with human entries in the primate (PRI) division of GenBank, combines these with human ESTs and creates clusters of sequences that share virtually identical 3' untranslated regions (3'UTRs). In this manner, the nearly 500000 human ESTs in dbEST have been reduced 10-fold to 50000 sequence clusters, each of which may be considered as representing a single human gene.

Thus, UniGene, in contrast to methods used in connection with the sequence *domain* analysis referred to in Office Action paragraph 7, actually endeavors to partition GenBank sequences into disjoint subsets, i.e. to perform a real clustering (of sequences, based on sequence similarity). One might therefore effect a reduction of sequence database size along the lines that were suggested in Andrade et al., page 409, column 1, para. 4, which is "to filter the databases for non-redundancy

at a level lower than that presently used in GeneQuiz", namely, by filtering out all GenBank sequences that do not lie within a specified distance from the centroids of each of the UniGene sequence clusters. However, to repeat the point made above, this hypothetical use of clustering still would be unrelated to the instant steps (a) and (b), because it does not pertain to the partitioning of a set of input genes as in the instant (a) and (b). Furthermore, this hypothetical use of clustering is not suggested by Andrade et al.

**8A.** On page 3, paragraph 8 of the Office Action, the Examiner refers to documents by Benson et al. and Bairoch et al., to expand on the point made on page 2, paragraph 6 of the Office Action. For reasons indicated below, Applicant disputes the Office Action's suggestion on page 3, paragraph 8 that "PROSITE provides documents directed to known protein families (Abstract etc. and pages 218-219)" and especially to the suggestion that what PROSITE provides is "as in the instant steps (c) and (d)".

**8B.** As indicated in paragraph 6 above, Applicant objects to the suggestion that GeneQuiz links to literature databases or to abstracts. The databases shown in Table 1 of Andrade et al (page 394), to which GeneQuiz links, such as Genbank and PROSITE, are databases of sequences (nucleic acid and protein). They are not literature databases, such as MEDLINE /Entrez PubMed which do contain literature abstracts. Furthermore, the GeneQuiz system that is disclosed by Andrade et al. does not make use of such literature abstracts, because abstracts do not in general embody annotations such as those illustrated in Table 3 (page 397) of Andrade et al. Consequently, such literature abstracts would not be used as described by Andrade et al., even if the GenBank and PROSITE sequence databases did actually contain literature or literature abstracts.

8C. The Benson et al. and Bairoch et al. papers that were attached to the Office Action confirm that Genbank and PROSITE, which are sequence databases used by GeneQuiz, are not literature databases and do not contain literature abstracts. Benson et al. (page 2, column 1, line 6) state that:

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, ... Bibliographic references are included along with links to MEDLINE unique identifiers for all published sequences.

By "published sequences" Benson et al. mean that researchers may not only be depositing a sequence into GenBank, but they may also be publishing a scientific manuscript in which the sequence is listed:

GenBank staff can usually assign an author an accession number within 1 working day of receipt. ... Authors have the right to request that their sequences be kept confidential until the time of publication. In those cases, authors are reminded to inform the database of the publication date in order to have a timely release of the data (page 4, column 2, line 11).

Thus, for each GenBank sequence entry, there is expected to be at most a single literature citation (not the actual literature and not a literature abstract), and that literature citation refers to a scientific publication (if any) in which the sequence appeared. Therefore, GenBank does not provide a means for associating a set of documents with each gene in a set of genes as expected in the instant step (c) because the sequences in GenBank (which may or may not correspond to genes) would be expected ordinarily to be associated with either zero or one document. This is contrary to the meaning of the term "set of documents" which implies at the very least two documents. Furthermore, GenBank does not provide a means for receiving the text of part or all of each of its documents as required in the instant step (d). (See below for a discussion of ENTREZ). Moreover, the instant step (c) refers to clusters of genes in the instant step (b), and there is nothing in GenBank or PROSITE that is concerned with the clustering of genes. As

indicated above, the clustering of genes (on the basis of sequence similarity) is found in UniGene, not GenBank, and this is unrelated to the clustering of the input to GeneQuiz, whereas Applicant's claims deal with the clustering of input genes based on properties that are unrelated to sequence similarity.

**8D.** Similarly, in Bairoch et al., which describes the PROSITE sequence database, the reference indicates that PROSITE contains only literature citations related to protein motifs (not actual literature and not literature abstracts) in its Fig. 1a (page 218), which may be what the Examiner is referring to in Office Action page 3, paragraph 8. PROSITE is a database of functional motifs, not of genes. The protein associated with a gene may or may not contain a functional motif, and if it does contains a functional motif, that motif may or may not be present in PROSITE, and there may be more than one functional motif within the protein corresponding to a gene. Therefore, PROSITE does not provide a means for associating a set of documents with a gene having no associated protein motif, as expected in the instant step (c), and PROSITE does not provide a means for receiving the text of part or all of each of its documents as required in the instant step (d). Moreover, the instant step (c) refers to the clusters of genes in the instant step (b), and there is no clustering of the input to GeneQuiz, whereas Applicant's claims deal with the clustering of input genes based on properties that are unrelated to sequence similarity.

**8E.** By referring on page 3, paragraph 8 of the Office Action to the ENTREZ system, the Examiner appears to be suggesting that a set of documents may be obtained as in the instant steps (c) and (d) through use of the ENTREZ system, in combination with literature citations or other items contained within the sequence databases GenBank or PROSITE. Applicant objects to this suggestion because this suggestion is not found in the teachings of Andrade et al. about the

GeneQuiz system, which makes no use of literature databases or literature abstracts, and therefore has no reason to make use of those components of the ENTREZ system that are intended to search literature databases.

8F. Although Applicant believes that it is incorrect to suggest that the GeneQuiz system makes use of literature abstracts as was indicated in paragraphs 6 and 8 of the Office Action, Andrade et al. did suggest improvements to the GeneQuiz system in which literature abstracts might be used (although not in a manner that anticipates or suggests the instant step (c) that was disclosed by Applicant, and not in a manner that involves GenBank, Prosite, or Entrez). On page 409, column 2, line 13 of Andrade et al., it is written that:

Likewise, the method of keyword analysis for functional classification can be extended to include wider sources of information. ... The possibility of direct extraction of keywords from bibliography databases, such as MEDLINE, is currently being explored (Andrade and Valencia, 1998).

The cited reference Andrade and Valencia (1998) then describes how “the method of keyword analysis for functional classification” that was described in Andrade et al. could be extended to include the extraction of keywords from bibliography databases such as MEDLINE, on its page 601, column 2, line 34 through page 602, column 1, line 10:

To obtain a representative set of words (and their abundances) in protein families, we selected a subset of distinct non-overlapping protein families. [NOTE: Andrade uses the term “protein families” in the same sense that Sonnhammer et al. uses the term “protein *domain* families”, such that a protein may in general be a member of more than one family]. These were taken from PDBSELECT ... . Protein families were taken from the HSSO database ..., with each family corresponding to one of the PDBSELECT proteins. To ensure that the proteins contained in each family perform only one function, we selected only those proteins with >40% of sequence similarity to the master sequence of the family. The set of abstracts corresponding to each of the families was assembled with the MEDLINE pointers in the corresponding SwissProt entry of each protein. Very small protein families were excluded, i.e., those with less than five proteins linked to MEDLINE. ... To collect the abstracts, we used the SRS system ..., which provides convenient access to MEDLINE through the WWW. For example, a search for the MEDLINE files containing a word beginning with ‘plastocyanin’ can be performed through [http://www.embl-heidelberg.de/srs/srcs? \[MEDLINE-AllText:plastocyanin\\*\]](http://www.embl-heidelberg.de/srs/srcs? [MEDLINE-AllText:plastocyanin*]).

The methods described in the above paragraph are not as in the instant step (c) because they pertain to a means for associating a set of documents with protein domain families, rather than a means for associating a set of documents with individual genes. Moreover, the instant step (c) refers to clusters of genes in the instant step (b), and there is no clustering of the input to GeneQuiz or anything suggested by the above paragraph, whereas Applicant's claims deal with the clustering of input genes based on properties that are unrelated to sequence similarity. However, the methods described in the above paragraph do provide a means for receiving text as in the instant step (d) because they make use of "the SRS system which provides convenient access to MEDLINE files through the WWW".

8G. Applicant is not representing in the instant step (c) that there no prior art for associating a set of documents with each gene in a set of genes, but rather that Applicant is unaware of prior art for associating a set of documents with each of the clusters obtained in the instant step (b). To appreciate the distinction, consider what was taught by Shatkey et al (2000). They considered a microarray experiment that had been performed by Spellman and colleagues. For that experiment, DNA corresponding to certain yeast genes had been placed on a microarray, and the expression of these genes was measured experimentally throughout the yeast cell cycle. Spellman and colleagues then grouped (clustered) these gene expression data. The method described by Shatkey et al did not use the experimental groupings of Spellman and colleagues, as would be required in the instant steps (a) and (b), but instead performed a *de novo* grouping of genes, based solely on an analysis of the literature about the genes. Thus, Shatkey et al write (on their page 8, column 2, line 52):

"The genes that are grouped as similar according to our method are compared with the ones grouped by functionality according to Spellman's table (parts of which are shown in Table 1)."

Consequently, Shatkay et al. teach away from the instant steps (a) and (b), and therefore teach away from the instant step (c), which requires that a set of documents be associated with pre-designated clusters in the instant step (b). In other words, Shatkay et al teach that when a clustering of genes is already available (as in the instant step b), then that clustering should be ignored except for purposes of manually comparing with results obtained independently by their method.

**8H.** This is not to say that Shatkay et al. do not associate a set of documents with individual genes, though, because such an association of documents with an individual gene is part of the method disclosed in their manuscript. That method involves the use of the PubMed literature database, which is well understood in the biomedical research community to be a component of the Entrez system that was referred to in the Office Action page 3, para. 8. For example, the web site for the sponsors of Entrez state the following at their web page

[www.ncbi.nlm.nih.gov/entrez/query/static/overview.html](http://www.ncbi.nlm.nih.gov/entrez/query/static/overview.html):

PubMed, available via the NCBI Entrez retrieval system, was developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), located at the National Institutes of Health (NIH). Entrez is a text-based search and retrieval system used at the NCBI for services including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others.

. The article by Shatkay et al (2000) explains (on page 3, column 1, line 55) that PubMed provides for literature search and retrieval by two methods, boolean query and similarity query (also known as “neighboring”):

The most widely used on-line source for gene-related abstracts is the PubMed database. An initial step in the search for relevant literature is the specification of a boolean query. ... This form of query suffers from several well-known deficiencies. ... An alternative to the boolean query paradigm is the use of similarity queries; the user provides a sample document that is relevant to the subject of interest and gets back other documents discussing the same subject matter. ...

In view of the deficiencies of the boolean query method, Shatkay et al then describe a technique that they developed using the alternate method (similarity queries), which required in their case to provide a sample document relevant to each gene that was used in the microarray experiment of Spellman and colleagues. The sample documents, which they termed “kernels”, were chosen as follows (page 8, column 1, line 45):

Out of about 800 genes found by Spellman et al to be cell-cycle regulated [in their gene expression microarray experiment], only 408 genes had curated PubMed references in the SGD [a yeast gene database], and our experiment concentrated on these 408 genes. For each of the genes, the oldest reference cited in SGD was chosen to be the kernel document corresponding to the gene.

8I. Thus, Shatkay et al associate an individual gene with a body of literature abstracts by first finding a citation for that gene (if possible) within a curated database about the genes under investigation, and then using the similarity query neighboring feature of PubMed to acquire a set of documents about that gene. As noted above, Shatkay et al teach that there are well-known deficiencies with any attempt to use of the other PubMed query method that might have been used, namely, the method of boolean queries. To illustrate those difficulties associated with boolean queries (text search strings), Applicant includes with this response to the Office Action a non-prior art paper by Chaussabel and Sher (2002), which attempts to use boolean queries and ultimately finds it necessary to manually edit or correct the unacceptably large number of errors that result from their attempt to automate the process:

The method requires articles related to each of the genes included in the analysis to be extracted. This is done by querying the Medline database through PubMed ... using appropriate search strings. The search for relevant literature for each individual gene is complicated by the fact that the same gene can have many different names associated with it and that the same name or abbreviation can have different meanings. A rapid scanning of the search results is useful for the identification and removal of inappropriate search strings (page 2, column 1, lines 27-53). ... Most search strings must be edited on a gene-by-gene basis, as a vast majority of publications do not adhere to the official nomenclature and gene names and abbreviations in use can differ from the

aliases provided by HGNC or lack specificity. Acronyms that contain only a few letters are particularly problematic and must be removed from the query... (page 14, column 2, line 19)

Accordingly, it appears to be impractical or impossible to use the boolean query feature of Entrez PubMed in this context without using an expert human as an integral part of the process of selecting a set of documents to correspond to a gene.

**9A.** On page 3, paragraph 9 of the Office Action, the Examiner writes that “GeneQuiz performs a lexical analysis of database annotations and decision criteria for functional assignments (text classification) (Abstracts etc.).” Applicant objects to the suggestion that what GeneQuiz does is as in the instant steps (e) and (f) because the primary decision criteria, by which GeneQuiz word annotations are assigned to query sequences, are totally unrelated to any lexical analysis; because the GeneQuiz lexical analysis does not involve any type of word weight-setting method (much less those implemented by the Rainbow program); because the GeneQuiz analysis is not being applied to any type of cluster; and because GeneQuiz is not being applied to documents (Abstracts or otherwise). Thus, GeneQuiz performs a transfer of pre-existing word and phrase annotations based on numerical criteria that are totally independent of the text itself (e.g., how similar the query sequence is to a database sequence according to BLAST and FASTA, and which sequence database contains the homologous sequence that is similar to the query sequence.), as described in Andrade page 398. Only part of this process of transferring pre-existing word and phrase annotations involves a lexical analysis (Andrade et al, p. 398, Col. 2), which is to remove words and phrases if those words or phrases occur in a stop-list of unacceptable words (like “hypothetical” or “unknown”) or if the words contain numbers or have fewer than five letters. Even this “lexical part” of the overall process of transferring annotations is unlike step (e) because the GeneQuiz lexical analysis does not involve any type of word weight-setting method (much less

those implemented by the Rainbow program), because it is not being applied to any type of cluster, and because it is not being applied to documents.

**9B.** On page 3, paragraph 9 of the Office Action, the Examiner writes that “GeneQuiz performs general functional class analysis based on the generation of a dictionary that associates keywords characteristics of a sequence with a set of functional class (set of sequences). From a training set, keywords are extracted and each is scored by the number of times that appears in a functional class (weighting) and assignment of a new sequence to a class is by lookup of the keyword for that sequence in a dictionary to determine the most frequently associated class.” Applicant objects to the suggestion that what GeneQuiz does in this regard is as in the instant steps (e), because the instant invention does not classify anything, does not make use of any type of training set, and does not involve (much less seek to assign the class for) any new (query) sequence. In the instant invention, genes are assigned to clusters in the instant step (b) through a partitioning of the set of genes in the instant step (a). There is nothing in the claims of the instant invention to suggest that the purpose of that assignment is for purposes of classifying some other entity, much less a classification on the basis of some new input entity on the basis of pre-existing text.

**9C.** On page 3, paragraph 9 of the Office Action, the Examiner writes that “GeneQuiz provides a means of sorting via the GQreason module by reliability values and categories (page 398, column 2, 21-25 and Table 5) and storing (page 393, column 1, lines 14-17) as in the instant steps (e) and (f).” Applicant objects that the sorting that is done by GeneQuiz (page 398, column 2, 21-25 and Table 5) is based on numerical criteria that are totally independent of text analysis (e.g., how similar a query sequence is to a database sequence according to BLAST and FASTA,

and which sequence database contains the homologous sequence that is similar to the query sequence), whereas Applicant's invention is unrelated to sequence analysis. Applicant also objects that the storing referred to on page 393, column 1, lines 14-17 refers to the storage of matter unrelated to the instant invention. That text reads "The GQsearch module applies a variety of sequence analysis tools to the query sequence, parsing, and storing the results in a common format for subsequent processing stages." The objection is that Applicant's invention does not involve a query sequence and does not apply sequence analysis tools to a query sequence (such as those described in Table2, page 395), and therefore the storage referred to in the instant step (f) does not pertain to such matters.

10. On page 4, paragraph 10 of the Office Action, the Examiner writes that "However, Andrade et al. does disclose the limitations of word weight-setting methods implemented by the computer program Rainbow". This sentence is obscure because the Andrade et al paper does not refer to the computer program Rainbow, and Applicant is unable to locate within Andrade et al any discussion that mentions the term "word weight-setting methods." If by "word weight-setting methods" the Examiner means that Andrade et al use the sequence similarity numerical scores from BLAST and FASTA, along with a biasing scheme favoring certain sequence databases (SWISS-PROT > PIR, etc.) in order to rank-order the transfer of homologue annotations to a previously un-annotated query sequence (page 398, col. 2), then Applicant objects because the computer Rainbow (and the instant invention) is totally unrelated to these methods for assigning numerical weights to words. This is because Rainbow performs an *ab initio* calculation of word-weights using a text corpus, rather than a transfer of pre-existing word and phrase annotations based on numerical criteria that are independent of the text itself (e.g., how similar two sequences are according to BLAST and FASTA, and which sequence database contains the sequence that is

similar to the query sequence.). In fact, GeneQuiz does not even make use what one would ordinarily consider to be a text corpus (see Declaration of David Rigney, paragraph 10).

11. On page 4, paragraph 11 of the Office Action, the Examiner writes that "McCallum discloses a program, Rainbow, for performing text and document classification (Rainbow, pages 1-11) as in the instant step (e)". Applicant calls the Examiner's attention to the text on page 1 of this disclosure that states that "Several of the examples also assume that you have downloaded the 20 newsgroups dataset, unpacked it in your home directory, and therefore that its files are available in the directory ~/20\_newsgroups" and to the text on page 8 of the disclosure that actually lists the most of the newsgroups dataset (alt.atheism, comp.graphics, etc). As indicated in the accompanying **Declaration of David R. Rigney**, the number of words in the text corpus corresponding to each of the sample classes is as follows, as an indication of the size of the text corpus with which the program Rainbow is expected to work:

<u>Class Name</u>	<u>Number of Words in the Text Corpus Corresponding to that Class</u>
alt.atheism	354053
comp.graphics	278585
comp.os.ms-windows.misc	234915
comp.sys.ibm.pc.hardware	216999
comp.sys.mac.hardware	203473
comp.windows.x	305914
misc.forsale	164281
rec.autos	237731
rec.motorcycles	217896
rec.sport.baseball	249160
rec.sport.hockey	301787
sci.crypt	348884
sci.electronics	225887
sci.med	313044
sci.space	310385
soc.religion.christian	404170
talk.politics.guns	356830
talk.politics.mideast	523816
talk.politics.misc	436764
talk.religion.misc	362082

Although Andrade et al do not discuss the size (in words) of the annotations that are associated with the sequences and dictionary of their method, it appears from the example shown in Table 3 (page 397), that the size (in words) is several orders of magnitude smaller than that of the text corpus expected by Rainbow. Therefore, there is nothing in Andrade et al that would suggest combining its methods with that of McCallum (1998).

**12A.** On page 4, paragraph 12 of the Office Action, the Examiner writes that “An artisan of ordinary skill in the art at the time of the instant invention would have been motivated by Andrade et al. to improve GeneQuiz to provide better sensitivity during searches and an increase in speed and accuracy (page 409, column 1, 42-51) [and] therefore, make the automated genome sequence analysis and annotation system wherein the text and document classification is performed by Rainbow as taught [by] McCallum.” Applicant objects to this statement for several reasons. First, if “the automated genome sequence analysis and annotation system” refers to Applicant’s invention, Applicant objects because Applicant’s system does not analyze sequences. Instead, it analyzes clusters of genes that have been identified on the basis of the similarity of their gene expression in a microarray experiment, which would generally be unrelated to any inter-cluster or extra-cluster similarity of the genes’ sequences and which is not analyzed in the instant invention. Furthermore, for the reasons explained above, a gene and a sequence are not synonymous. Second, the invention that was disclosed by Applicant does not perform document classification. This is because for the claims of this invention, there is no step in which a document is presented to the claimed invention wherein Rainbow is asked to classify the document. The Examiner may be confusing the present invention with claims that had been withdrawn (e.g., withdrawn Claim 2). The instant steps (a), (b), and (c) pertain to the assignment of documents to a cluster, which is

not a classification of documents in the ordinary sense of the word "classify", and this assignment does not make use of the Rainbow software. Third, if "classify text" means to find key words or phrases, then the instant invention does so, on the basis of which words or phrases occur frequently in text about each cluster AND simultaneously do not appear frequently in text about other clusters. However, this function is only part of what Rainbow does, which is a preliminary or adjunct to its main function of classifying the text within documents. Rainbow's classification function is not used in the present invention, although it is used for the withdrawn claims.

**12B.** Fourth, Applicant disputes whether an artisan of ordinary skill in the art at the time of the instant invention would have been able to simultaneously understand the teachings of Andrade and of McCallum, much less combine them. As described in the accompanying **Declaration of David Rigney**, Applicant believes that an artisan of ordinary skill at the time of the present invention is typified by a group of individuals called the "Boston Area Molecular Biology Computer Types", of which David Rigney was formerly a member. As described in the web site for that group, [genetics.mgh.harvard.edu/bambct/bambct-mission.html](http://genetics.mgh.harvard.edu/bambct/bambct-mission.html), most of its members provide hardware and software support to university molecular biology departments or BioTech companies (in the Boston area). The group started out with most people running the GCG Wisconsin package of sequence analysis software but some run similar programs (like DNA\*) or specialize in sub-areas of molecular biology computing. About half of the group have doctorates in a technical subject and the other half have bachelors or masters degrees in a technical subject that requires some practical knowledge of computer hardware and software (computer science, physics, chemistry, engineering). According to this proposed typecasting of "an artisan of ordinary skill", the artisan would ordinarily be able to write simple computer programs (scripts) for pre-existing software (like the GCG Wisconsin package) but not necessarily be able to write whole,

compiled computer programs that require the design and implementation of a new algorithm. The common knowledge among such a group of artisans would be the subject matter of the GCG Wisconsin Package, which is summarized in Attachments 2 and 3 for the accompanying **Declaration of David Rigney**. Those attachments demonstrate that many of the topics discussed in Andrade et al (1999) are also to be found in the GCG Wisconsin package, such as BLAST, FASTA, GenBank, and Pfam. However, there is nothing in the GCG package dealing with microarrays, and there is nothing in the GCG package that is concerned with machine learning or of text analysis. Therefore, the latter topics would be considered to pertain to specialized sub-areas of molecular biology computing, not well understood by the artisan of ordinary skill.

**12C.** Andrade et al. disclose a system for automatically annotating sequences. The invention that was disclosed by Applicant, on the other hand, is not an automated sequence analysis and annotation system. It is instead an automated microarray cluster analysis and annotation system. If by coincidence there were some sequence similarity among genes in a microarray cluster, this fact would not be known to or used by the invention disclosed by Applicant. The only suggestion in Andrade et al. that is even remotely related to the field in which microarrays are used (gene expression studies, cell physiology) appears on page 409, column 2, line 28, where it is stated that:

Radical changes to the architecture of the GeneQuiz system may be envisaged. In the current version, the analysis of a related batch of sequences from a complete genome makes no use of any genome-, cell-, or organism-level information: each sequence analysis is independent. However, the sequences are connected ... (ii) systematically by their cooperative participation in the same cellular and/or organismal entity (e.g. metabolic pathways or signalling cascades, either specific to particular tissue types or general to organism as a whole). The initial conclusions inferred in one pass of analysis could be appraised in the context of such relationships and then fed back into subsequent refinement cycles, maybe modifying previous conclusions.

The clustering of microarray data that is contemplated by Applicant's invention is intended precisely to *discover* heretofore unknown metabolic pathways and the like, for a particular cell type, whereas the text cited above supposes that such cell physiological pathways or relationships are *already known* and are to be used in some unspecified manner to appraise the conclusions produced by GeneQuiz concerning individual genes, and only after some unspecified "radical change" is made to the architecture of the GeneQuiz system. Therefore, considering that there is nothing in Andrade et al that is related to the use of microarrays, and considering that the artisan of ordinary skill described in the **Declaration of David Rigney** would not in general be technically familiar with the methods of gene expression analysis (including the use of microarrays), there is nothing in Andrade et al to suggest Applicant's invention to the artisan of ordinary skill.

**12D.** As described in the accompanying **Declaration of David Rigney**, if the artisan of ordinary skill were presented with Andrade et al (1999), the artisan might follow the suggestion on its page 409, column 2, lines 24-27 to study the referenced article Andrade and Valencia (1998) in an effort to "explore the direct extraction of keywords from bibliography databases". After reading the article Andrade and Valencia (1998) and then being made aware of McCallum (1998), the artisan might realize that the methods described by McCallum (1998) may be applied to the problem described by Andrade and Valencia (1998), because the Andrade and Valencia method "estimates the significance of words by comparing the abundance of words in a given set of abstracts related to a protein family with their abundance and distribution in a background set of abstracts associated to a wide range of different proteins". However, the actual methods described by Andrade and Valencia (which are ad hoc) are different from the methods described by McCallum (1998), so the ordinary artisan, who is not experienced in the subject of machine

learning, would then be expected to seek background information beyond that described in McCallum (1998), which states on its page 1 that "This documentation is intended as a brief tutorial for using rainbow, version 0.9 or later. It is not complete documentation. It is not a tutorial on the source code". Because there are no background references within the Rainbow software itself or its documentation, the artisan may after substantial effort discover that some of the background information needed to understand and use Rainbow is located in Chapter 6 of the disclosed prior art Mitchell (1997) and references cited therein, concerning Rainbow itself, the Naive Bayes algorithm, as well as the several alternative algorithms implemented by Rainbow. The artisan must then be able to actually install Rainbow on a computer. Considering that Rainbow is not commercial software, troubleshooting may be required, so that the artisan would most likely have to have experience in computer systems management/programming – which may also be beyond the skills of the ordinary artisan.

12E. The artisan would then have to be able to combine the teachings of McCallum with that of Andrade and Valencia (1998), in order to provide a combined method that "estimates the significance of words by comparing the abundance of words in a given set of abstracts related to a protein family with their abundance and distribution in a background set of abstracts associated to a wide range of different proteins". Considering that the software described by Andrade and Valencia (1998) is available only as a server at the internet address columbia.ebi.ac.uk:8765/andrade/abx and not as actual source code, the artisan would have to be able to design and write his or her own compiled software to do so, including software for actually downloading and constructing a corpus of text from MEDLINE/PubMed. It is doubtful that an ordinary artisan, considered to be a member of the "Boston Area Molecular Biology

Computer Types", would be able to do write such software because the ordinary artisan can only write scripts.

12F. Even if that artisan were able to write such software, it is very unlikely that artisan would consider using such software as described in steps of the instant invention. This is because the only suggested use for such software in Andrade et al (1999) is in connection with the method of keyword analysis for functional classification, which, as stated on page 409, is to be extended to include wider sources of information such as bibliography databases. It is also stated on page 409 that the purpose of that extension is to increase the accuracy of classification, but there is no clear suggestion as to how that increase in accuracy is to be accomplished, other than that it somehow involve the extraction of keywords from bibliography databases. In the current method of keyword analysis for functional classification (see page 397), assignment of a new sequence to a class is by look-up, of the keywords associated with annotated sequence homologues of a query sequence, in a dictionary of class-associated keywords, to determine the class having keywords that most frequently match the homologue keywords, which is then chosen as the appropriate class ( see page 397). By analogy with the objectives stated in Andrade and Valencia (1998) that the method "estimates the significance of words by comparing the abundance of words in a given set of abstracts related to a [something] with their abundance and distribution in a background set of abstracts associated to a wide range of different [something]", the most logical application of the method of Andrade and Valencia (1998) to the GeneQuiz software in Andrade et al (1999) would be that [something] in the quotation above is to be "functional class". That is to say, if the ordinary artisan were to reach this level of understanding, the artisan would be tempted to replace the dictionary in Andrade et al (1999) with a dictionary constructed through use of the methods of Andrade and Valencia (1998) combined with McCallum (1999).

**12G.** However, it is not obvious that this understanding could then lead to steps in the present invention because by “functional class” Andrade et al mean classes whose function are already understood (see Table 4, page 398), whereas the “clusters” contemplated in steps (b) and (c) of the instant invention refer to groups of genes whose function are not only not understood, but that may even be spurious in the sense that they might arise through noise in the data of a microarray experiment. In fact, the purpose of the keyword analysis of the instant invention is generate keywords to assist in deciding whether the clusters are even meaningful. The purpose has nothing to do with any sequence analysis, which is the intended use of GeneQuiz. It would not be obvious to an expert in sequence analysis, microarrays, and machine learning to convert the functional sequence classes of GeneQuiz into the microarray clusters of the instant invention, so it would be even less obvious to an ordinary artisan, who would not be expected to be technically familiar with microarrays or machine learning.

**12H.** On page 4, paragraph 12 of the Office Action, the Examiner writes that “Therefore, it would have been obvious to one having ordinary skill in the art at the time of the invention was made to make an automated genome sequence analysis and annotation system to have a computer program, Rainbow, to perform the text and document classification, as taught [by] Andrade and McCallum. Applicant disputes this statement for the reasons described in the previous paragraphs, for example, because Applicant’s instant invention is not an automated genome sequence analysis and annotation system; because the size of the text expected by the computer program Rainbow is several orders of magnitude larger than that used by Andrade et al; because the combination of Andrade et al and McCallum needed to arrive at Applicant’s invention is not suggested in Andrade et al and McCallum; and because the artisan having

ordinary skill in the art would not be able to combine Andrade and McCallum to arrive at Applicant's invention.

THE DIFFERENCES BETWEEN THE PRIOR ART AND THE CLAIMS IN ISSUE

13A. Differences between the prior art and the claims in issue were discussed in the previous paragraphs. What follows is a summary of some of those differences, listed according to the paragraph in which a difference was identified.

13B. GeneQuiz links to sequence databases and does not link to literature databases. The instant invention links to literature databases and does not link to sequence databases (see paragraph 6 above).

13C. GeneQuiz does not analyze groups of input sequences, except in the trivial sense that it permits its user to make independent, sequential sequence queries. The instant invention performs an analysis simultaneously, involving groups of subsets of genes (see paragraph 7A above). Accordingly, by the method of Andrade et al., word or phrase annotations that occur among pre-existing annotations for sequences that are similar to the query sequence are considered, irrespective of whether those words or phrases also occur among the annotations for sequences that are *dis*-similar to the query sequence; *whereas* Applicant's invention teaches that words or phrases that occur frequently within the literature about a particular microarray cluster may be given a large or small numerical weighting, depending on whether those words or phrases also occur frequently among the words or phrases that occur frequently within the literature about different microarray clusters.

13D. GeneQuiz is concerned with protein domain families. The instant invention is concerned with genes, irrespective of whether the gene is associated with a protein domain (see paragraph 7E above).

13E. GeneQuiz does not attempt to partition objects into disjoint sets. The instant invention partitions genes into disjoint sets, namely, clusters (see paragraph 7E above).

13F. GeneQuiz attempts to match an input sequence with sequences in sequence databases using utilities like BLAST. The instant invention does not perform sequence matching (see paragraph 7E above).

13G. GeneQuiz endeavors to transfer pre-existing annotations based on numerical criteria involving sequence similarity and usage of particular sequence databases. The instant invention creates annotations ab initio, using criteria that are unrelated to sequence similarity or sequence databases (see paragraph 9A above).

13H. Lexical analysis within GeneQuiz does not involve any type of word weight-setting method. Lexical analysis within the instant invention involves numerical word weight-setting (see paragraph 9A above).

13I. GeneQuiz attempts to classify individual objects. In the instant invention, genes are simultaneously assigned to clusters, which is not classification in the same sense of the word "classify" (see paragraph 9B above).

13J. GeneQuiz does not use a text corpus of the size with which the program Rainbow is expected to operate. The instant invention uses Rainbow to operate on a text corpus (see paragraph 11 above).

**13K.** GeneQuiz makes no use of cellular, tissue, or organismal information. It is concerned only with sequence analysis. The instant invention makes use of cellular and tissue information through the use of microarray expression data (see paragraph 12C above).

**13L.** GeneQuiz is concerned with functions that are already understood (protein domain families). The instant invention is concerned with functions that are not only not understood, but that may even be spurious, namely, unknown functions of microarray clusters (see paragraph 12G above).

**13M.** Therefore, Applicant and Andrade et al. teach fundamentally methods of annotation, except that within the whole method of Andrade et al. there is a sub-method that is concerned with the construction of a dictionary, the use of which cannot be considered to be independent of Andrade's whole method.

**MODIFICATION OF THE APPLIED REFERENCES NECESSARY TO ARRIVE AT THE CLAIMED SUBJECT MATTER**

**14A.** The only part of GeneQuiz that might be suitable for modification to arrive at the claimed subject matter is the following section that is concerned with the construction of a dictionary, as described on page 397, column 1, line 6 and continuing to page 398, column 1, line 3:

The method used by GeneQuiz for general functional classification is based on the generation of a dictionary that associates keywords characteristic of a sequence with a set of functional classes. ... GeneQuiz currently works with 14 classes of cellular function ... (Table 4) ..., although the following algorithm can be applied to any classification scheme or number of classes. The generation of a dictionary starts with an initial comprehensive training set of example proteins classified by a human expert. For every one of those proteins, their corresponding keywords are extracted [NOTE: I assume the extraction to be from pre-existing annotations in the corresponding entries of the databases in Table 1, p. 394] and each is scored by the number of times that it appears in a functional class. A filtering procedure is applied to eliminate those keywords with no functional meaning ... and those that are present in just one sequence. Each one of the resulting set of keywords is assigned uniquely to a functional class if no less than 85% of its occurrences belong to that class. Assignment of a new sequence to a class is by look-up of the keywords for that sequence [NOTE: I assume the keywords for the new sequence to be from pre-

existing annotations in the corresponding entries of the databases in Table 1, p. 394] in the dictionary to determine the most frequently associated class, which is then chosen. Iterative application of the assignment process to all sequences in a sequence database yields a new set of keyword/class associations that can be used to generate a more extensive dictionary with an increase in classification quality (Tamames et al., 1998).

This method is described in only a little more detail in Tamames et al (1998), which is a two page application note.

**14B.** If you take the method described in the text cited above out of the context of everything else that is done by GeneQuiz (*which violates the principle that the reference must be considered as a whole, and which violates the principle that the reference must be viewed without the benefit of impermissible hindsight vision afforded by the claimed invention*);

**14C.** AND if you do not require classifications like those shown in Table 4 of Andrade et al. to correspond to recognizable functions, but instead allow the clustering of microarray data to generate nameless classifications (*which is away from the teaching of Andrade that the classifications should be meaningful and which introduces a reasonable expectation of failure because the microarray cluster are of uncertain functional meaningfulness by virtue of noise in the original microarray data*);

**14D.** AND if instead of using a human expert to pick proteins (or their corresponding genes) to correspond to a training set for each of the classification classes, you use the genes assigned automatically to each of the clusters of the microarray data (*which is away from the teaching of Andrade that the classification exemplars should be meaningful and which introduces a reasonable expectation of failure because the genes defining the microarray cluster are of uncertain functional relatedness by virtue of noise in the original microarray data*);

;

**14E. AND** if you replace the above-mentioned filtering procedure and ad hoc 85% preferential occurrence rule with methods used in the Rainbow system described by McCallum (1998), such as the Baysean method (*knowledge of which may not have been generally available to one of ordinary skill in the art*);

**14F. AND** if you decide to retain use of a tiny (on the order of 100 words per class) text corpus consisting of annotations taken from sequence databases like those shown in Table 3 of Andrade et al., (*which is likely to be make the method fail considering that the methods described by McCallum are ordinarily performed using a large text corpus, typically on the order of 100, 000 words per class as described in the Declaration of David R. Rigney*);

**14G. OR** if you decide to replace the use of sequence database annotations with the use of a text corpus about genes from Pubmed/MEDLINE abstracts as disclosed by Shatkay et al (2000)), which is more likely to work with the methods described by McCallum (*but which still has an uncertain likelihood of success considering that the “kernel” documents for the genes in question would have to occur in some curated database, and would have to correspond to a sufficient number of PubMed/MEDLINE abstracts to generate a large enough text corpus with which the methods of McCallum would be expected to work*);

**14H. OR** if you were to independently invent the method for generating a text corpus that is disclosed in the instant patent application (*which is highly unlikely considering that apparently nobody else has done so to date, and which violates the principle that the*

*reference must be viewed without the benefit of impermissible hindsight vision afforded by the claimed invention);*

**14I. AND** if you abandon your intention to use the keywords generated for each class to assign a new sequence to a class by look-up of the keywords for that sequence, and thereby abandon the iterative use of the method with sequence database annotations to achieve acceptable classification quality (*which is fundamentally contrary to the objectives and teachings of Andrade et al., which violates the principle that the reference must be considered as a whole, and which violates the principle that the reference must be viewed without the benefit of impermissible hindsight vision afforded by the claimed invention);*

**14J. THEN**, you would have arrived substantially at the invention that was disclosed in the instant patent application.

#### CONCLUSION

**15A.** To establish a *prima facie* case of obviousness, three basic criteria must be met. First, there must be some suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to modify the reference or to combine reference teachings. However, as described above, lack of motivation within the references themselves, and the knowledge generally available to one of ordinary skill in the art, would not have motivated the artisan to combine the reference teachings. Second, there must be a reasonable expectation of success. However, there would not be a reasonable expectation of success for reasons described in the previous section. Finally, the prior art reference (or references when combined) must teach or suggest all the claim limitations. The teaching or suggestion to make the

claimed combination and the reasonable expectation of success must both be found in the prior art and not based on applicant's disclosure. However, as described in the previous section, the convoluted modifications that would have to be made to a particular section of the Andrade et al reference are such that the prior art reference, taken as a whole, do not teach all the claim limitations, without impermissible hindsight. Consequently, Applicant believes that the instant invention is patentable over Andrade taken with McCallum.

**15B.** Applicant requests pursuant to MPEP 707.07(j) that the Examiner draft one or more suitable claims for the applicant, if the Examiner finds patentable subject matter disclosed in this application, but feels that Applicant's present claims are not entirely suitable. Applicant requests in particular that the Examiner consider the drafting of one or more claims describing particular embodiments of the general Claim 1, along the following lines:

A system, method, or computer program product of Claim 1, wherein:

- (a) the means for identifying a set of genes is to acquire the list of accession numbers corresponding to spots on a nucleic acid microarray, then identify the UniGene number with which each accession number is associated;
- (b) the means for partitioning the set of genes in (a) is to acquire a list of clusters corresponding to the co-expression of those genes' mRNA, acquired by the system from a source that is external to the system;
- (c) the means for associating a set of documents with each gene in the set of genes in (a), wherein database linkages are traced from Unigene number to LocusLink number to Omim number to Literature UIDs contained within a Web page for the corresponding Omim number;
- (d) the means for receiving text for documents in (c), wherein the text corresponding to each Literature UID is downloaded from a PubMed database of the U.S. National Institutes of Health;

(e) the means for assigning numerical weights to words or phrases contained in the text in (d), wherein the text in (d) is partitioned according to its association with clusters as provided in (b) and (c), followed by the application to the words and phrases in the partitioned text documents, of the "print-word-weights" option of the computer program Rainbow, operating with its default parameters such as use of the Naive Bayes method;

(f) the means for sorting, storing, and displaying of the words and phrases of (e), wherein said words and phrases are written to a computer file that may be read by the User, said file containing adjacent columns of words & phrases and their associated numerical weights.

15C. Applicant respectfully requests that a timely Notice of Allowance be issued in this case because the response was filed within the period of time provided by 37 CFR 1.136(a).

Respectfully submitted,

David R Rigney

David R. Rigney, Inventor  
GENETWORKS Inc.  
P.O. Box 33296  
Austin TX 78764-0296  
Tel. 512-445-7301  
drigney@genetworks.com

**CERTIFICATE OF EXPRESS MAIL UNDER 37 C.F.R. 1.10**

I hereby certify that this paper is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated below and is addressed to Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450

Printed Name: David R. Rigney

Date of Deposit: July 7, 2004

Signature: David R Rigney

Express Mail Label No. ER 826633934 US